Введение. Задачи информационного поиска

Лекция 1

Максим Бусел,
busel.maksim@yandex.ru

БГУ ФПМИ, 2018

О курсе

- Курс на AnyTask: https://anytask.org/course/299
- Инвайт: 1YPQrz8
- все вопросы/предложения/пожелания присылать на busel.maksim@yandex.ru
- требования на зачет:
  1. сдать домашние задания
  2. набрать 60-70% от максимума

О курсе

Содержание

Литература

К. Маннинг, П. Рагхаван, Х. Шютце.
Введение в информационный поиск.
М.: Вильямс, 2014.

S. Buettcher, C. Clarke, G. Cormack.
Information Retrieval: Implementing and Evaluating Search Engines.
Massachusets Institute of Technology, 2010.

B. Croft, D. Metzler, T. Strohman.
Search Engines: Information Retrieval in Practice.
Addison Wesley, 2009.

План

План

Определение

Информационный поиск (IR) - это процесс поиска неструктурированного материала (документа), удовлетворяющего информационные потребности, в некоторой коллекции.

Информационный поиск (IR) – это область научных исследований, ориентированная на изучение структуры, организации, хранения, поиска и извлечения информации.

Смежные дисциплины

- Машинное обучение
- Компьютерная лингвистика
- Дискретная математика
- Распределенные системы
- Психология
- ...

Объекты исследования

- Web страницы
- Электронные письма
- Книги
- Новости
- Профили людей
- Картинки
- Видео

- Информация в документах носит, чаще всего, неструктурированный характер, в отличии от, например, записей в базах данных.
- Текстовые документы могут содержать мультимедийную информацию.

Приложения информационного поиска

- Поиск
- Фильтрация
- Кластеризация и классификация
- Автоматическое реферирование
- Извлечение фактов
- Ответы на вопросы

Тенденции развития

1. Свежесть и актуальность информации
2. Региональность
3. "Мобильность"

Тенденции развития в веб-поиске

- Разнообразие результатов ('колдунщики', специальные результаты для отдельных классов запросов)
- Поиск информации $->$ решение информационной задачи пользователя.

Информационные задачи пользователей

Выделяют следующие классы запросов:

1. Информационные в открытой/закрытой форме ('сколько весит африканский слон', 'метод Ньютона', 'дзэн')
2. Навигационные ('tut by', 'Приорбанк')
3. Транзакционные ('купить билет на park live 2016', 'скачать introduction to information retrieval')

План

Классификация IR по масштабу

1. Глобальный (Web Search)
2. Персональный (Desktop Search)
3. Корпоративный (Enterprise Search)
4. Предметно-ориентированный (Domain-specific Search)

Web Search

- Запрос (search query) формулируется в виде короткого текстового сообщения.
- Коллекция составляет порядка 10^{11} документов.

Web Search

- Запрос (search query) формулируется в виде короткого текстового сообщения.
- Коллекция составляет порядка 10^{11} документов.
- Результат возвращается пользователю в виде упорядоченного списка ссылок на документы и выжимок их текстов (snippets).
- Список результатов упорядочен по убыванию релевантности заданному запросу.

Web Search

- Коллекция является некоторой репрезентативной выборкой Интернета. Регулярно обновляется и пополняется (Web Crawling).

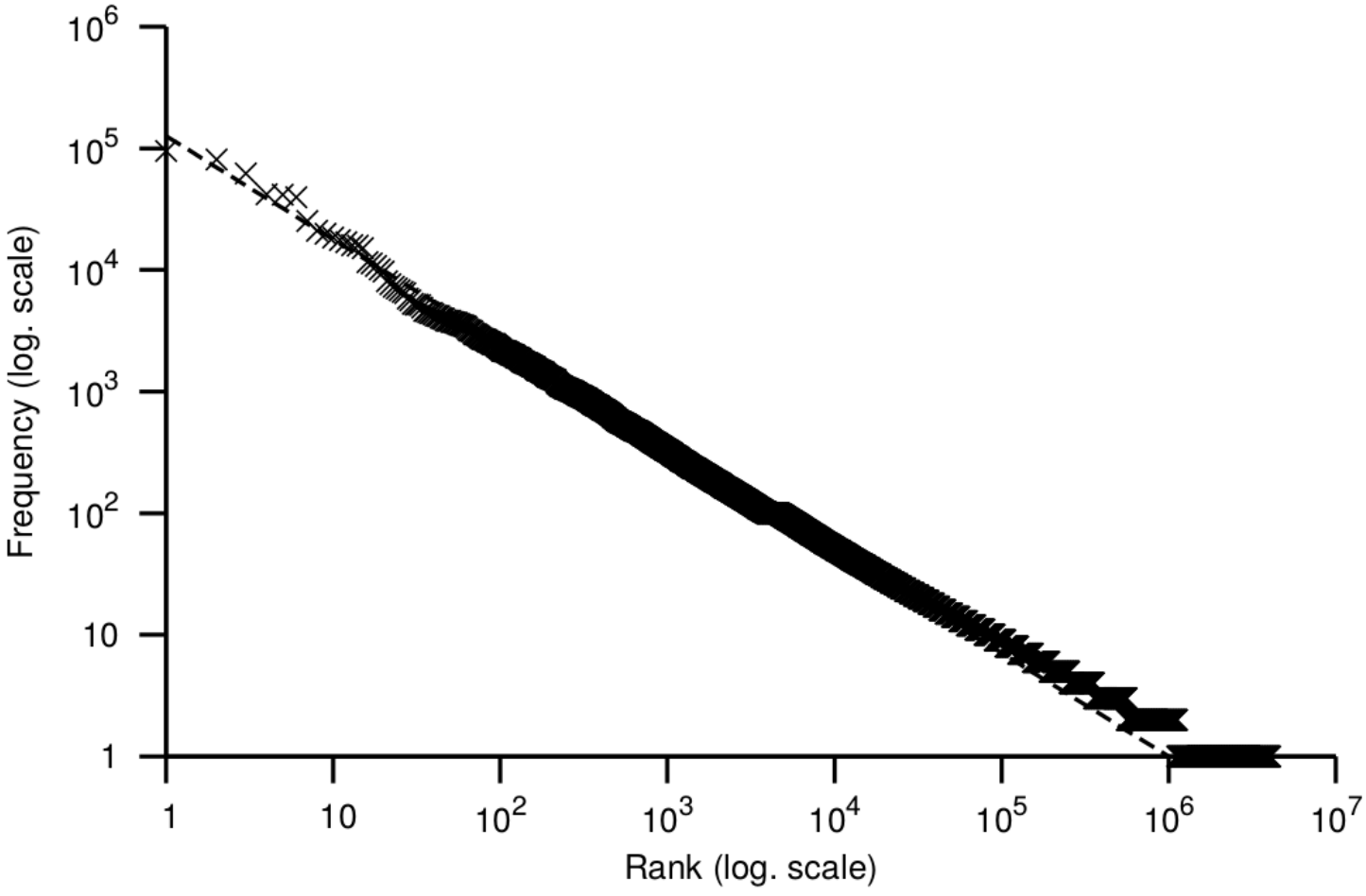
Web Search

- Коллекция является некоторой репрезентативной выборкой Интернета. Регулярно обновляется и пополняется (Web Crawling).
- Запросы в потоке распределены неравномерно (20% уникальных запросов $\sim$ 80% потока).

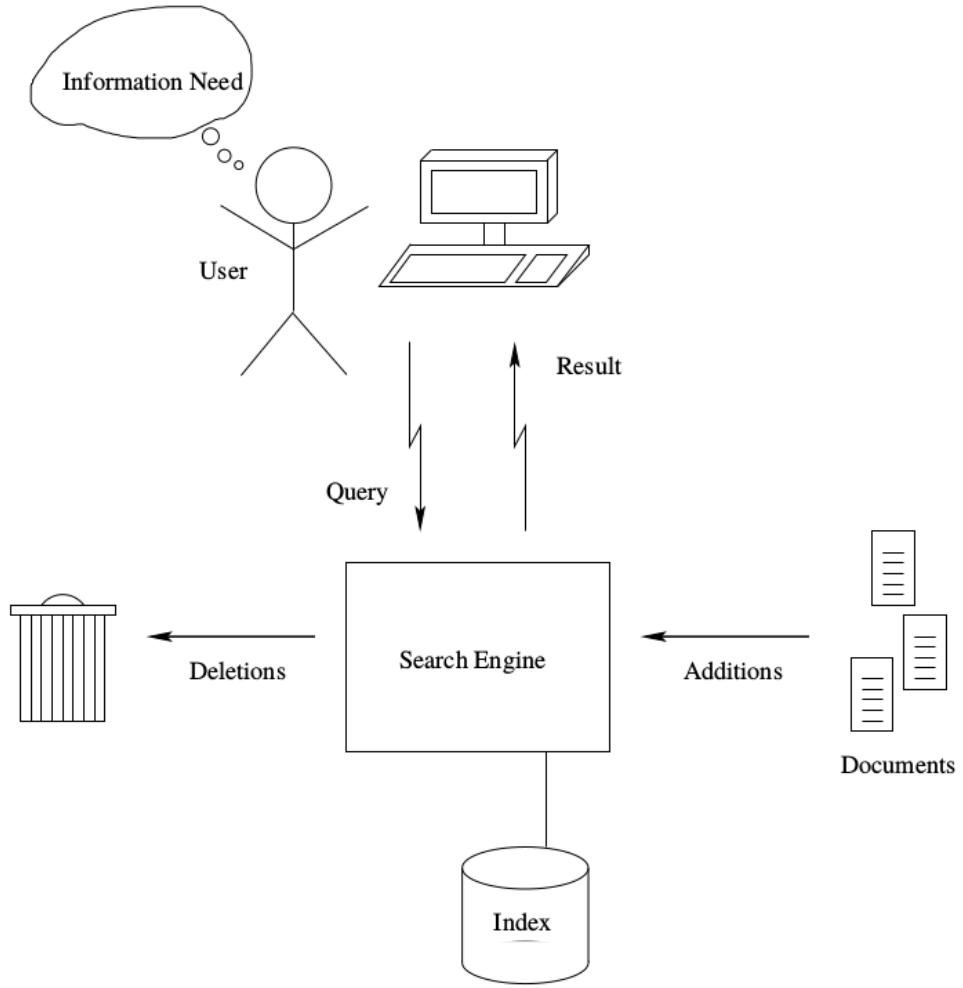
Web Search

- Коллекция является некоторой репрезентативной выборкой Интернета. Регулярно обновляется и пополняется (Web Crawling).
- Запросы в потоке распределены неравномерно (20% уникальных запросов $\sim$ 80% потока).
- Запросы задаются пользователями из различных географических регионов.

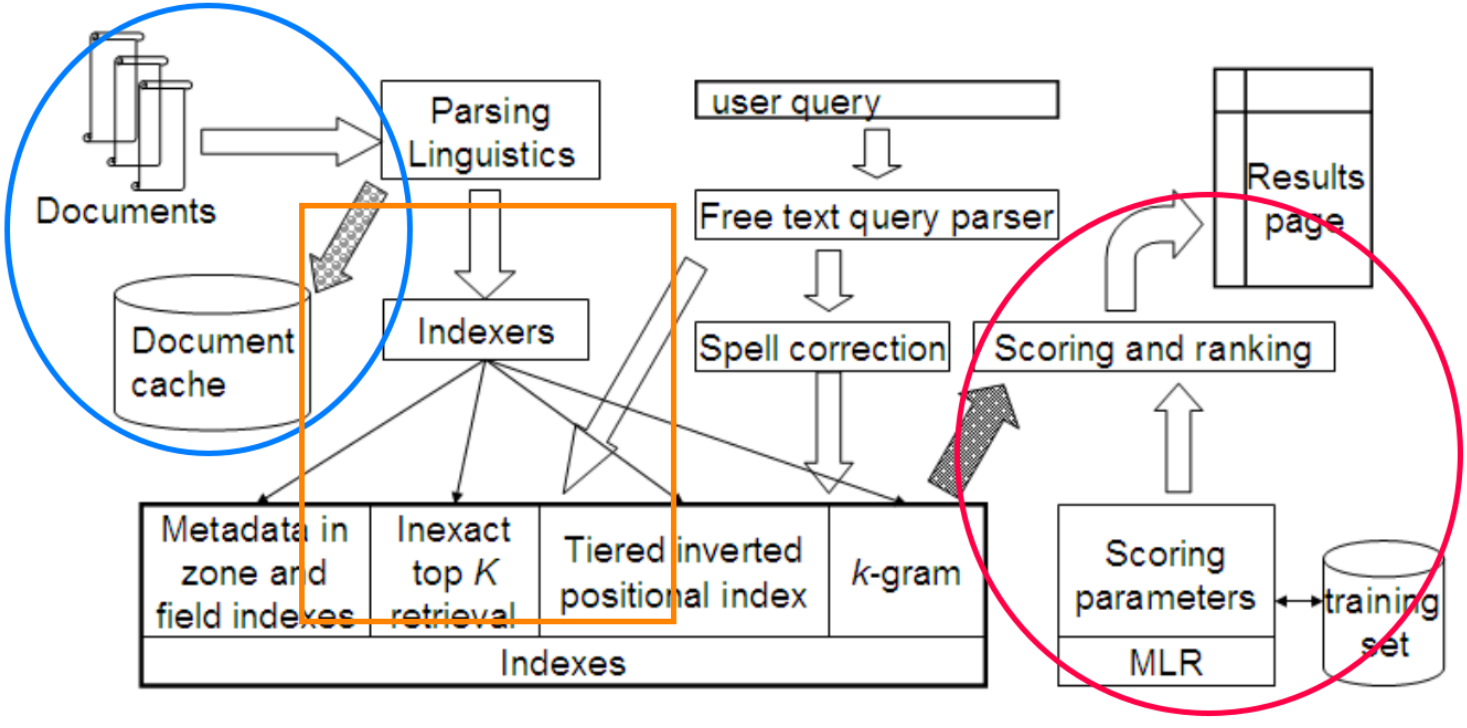
Распределение запросов в потоке



Упрощенная архитектура ИПС



Чуть менее упрощенная архитектура ИПС



Что должна уметь поисковая система?

1. Скачивать документы.
2. Работать с различными форматами и кодировками документов.
3. Понимать, когда обновляются документы, чтобы иметь актуальный индекс.
4. Находить дубликаты документов.
5. Обнаруживать спам.
6. Обрабатывать тексты запросов пользователей.
7. Ранжировать результаты.
8. Показывать сниппеты.

Probability Ranking Principle

Принцип вероятностного ранжирования (PRP)

Если результатом задания поискового запроса в ИПС является упорядоченный по убыванию вероятности релевантности список документов коллекции, то эффективность системы относительно пользователя максимальна.

Характеристики поисковой системы

1. Время ответа.
2. Скорость индексирования.
3. Качество ранжирования.
4. Полнота коллекции.

Пример: оценка неранжированных результатов поиска

Точность

$$P = \frac{D_{rel} \cap D_{retr}}{D_{retr}}$$

Полнота

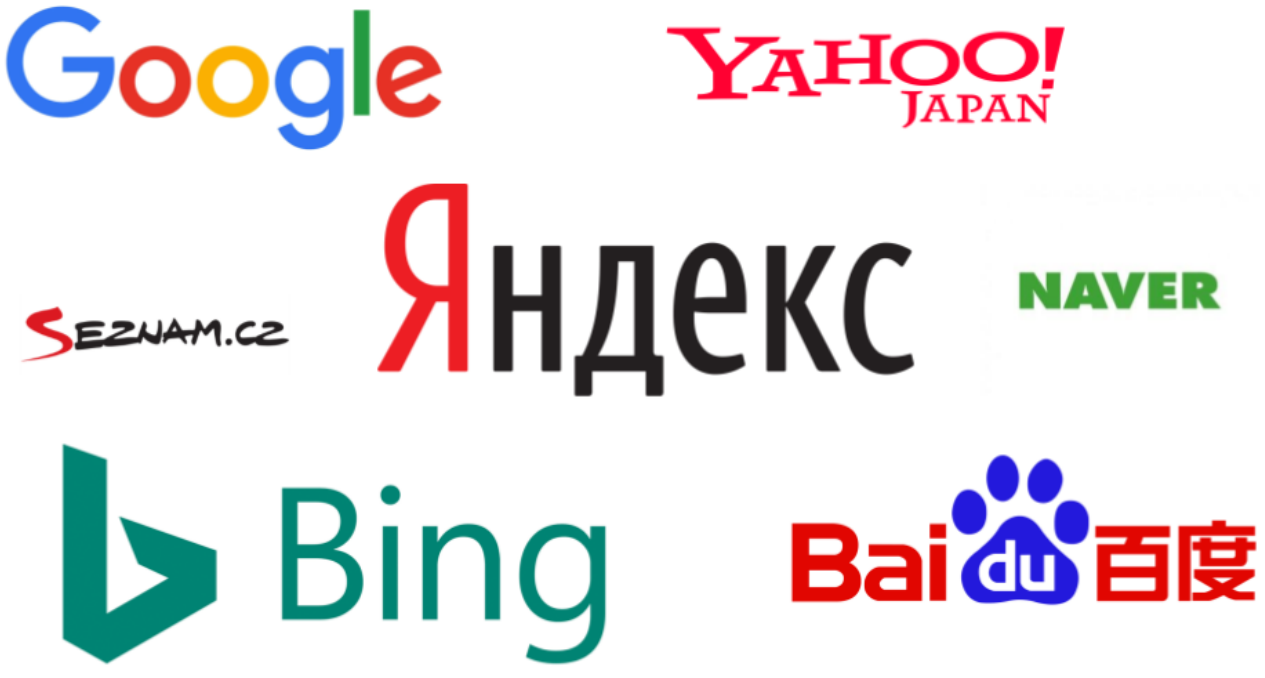
$$R = \frac{D_{rel} \cap D_{retr}}{D_{rel}}$$

F-мера

$$F = \frac{(1 + \beta^2)PR}{\beta^2 P + R}$$

Мировые лидеры

Поисковые системы, которыми пользуются более 20% пользователей в стране



Honorable mentions

Open-Source проекты

- Lucene (индексация, поиск)
- Lemur, Indri, Galago, RankLib (индексация, поиск, ранжирование, языковое моделирование)
- Sphinx (индексация, поиск)
- Wumpus
- Nutch (поисковый робот, интеграция с Hadoop)

https://en.wikipedia.org/wiki/List_of_search_engines

План

Конференции

- TREC
- WWW
- SIGIR
- ECIR
- WSDM
- KDD
- CIKM

Технологии

MapReduce (Google), Pregel (Google), GFS, browser toolbars, PageRank (Google), LambdaMART (Microsoft), DSSM (Microsoft), ...

Следующая лекция

Модели информационного поиска